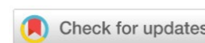
ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ
INFORMATION TECHNOLOGY, COMPUTER SCIENCE AND MANAGEMENT

УДК 57.087 *Научная статья*
https://doi.org/10.23947/2687-1653-2023-23-3-296-306

GATCGGenerator: новый генератор для создания квазислучайных нуклеотидных последовательностей

О.Ю. Кирьянова[1] ✉, **Р.Р. Гарафутдинов**[2], **И.М. Губайдуллин**[3], **А.В. Чемерис**[2]
[1] Уфимский государственный нефтяной технический университет, г. Уфа, Российская Федерация
[2] Институт биохимии и генетики — обособленное структурное подразделение Федерального государственного бюджетного научного учреждения «Уфимский федеральный исследовательский центр», г. Уфа, Российская Федерация
[3] Институт нефтехимии и катализа — обособленное структурное подразделение Федерального государственного бюджетного научного учреждения «Уфимский федеральный исследовательский центр», г. Уфа, Российская Федерация
✉ olga.kiryanova27@gmail.com

Аннотация
Введение. В последние десятилетия знания о ДНК все шире применяются для решения небиологических задач (вычисления с помощью ДНК, долговременное хранение информации). В первую очередь речь идет о случаях, когда необходимо подобрать искусственные нуклеотидные последовательности. Для их создания используются специальные программы. Однако существующие генераторы не учитывают физико-химические свойства ДНК и не позволяют получать последовательности с явно выраженной «небиологической» структурой. Фактически они генерируют последовательности, распределяя нуклеотиды случайным образом. Целью данной работы является создание генератора квазислучайных последовательностей с особой нуклеотидной структурой. Он должен учитывать некоторые физико-химические особенности нуклеотидных структур и будет задействован при хранении небиологической информации в ДНК.
Материалы и методы. Описано новое программное обеспечение GATCGGenerator для генерации квазислучайных последовательностей нуклеотидов. Оно предоставляется как SaaS (от англ. software as a service — программное обеспечение как услуга), что обеспечивает его доступность с разных устройств и платформ. Программа генерирует последовательности определенной структуры с учетом гуанин-цитозинового (GC) состава и содержания динуклеотидов. Представлена работа алгоритма новой программы. Требования к сгенерированным нуклеотидным последовательностям заданы с помощью чата в «Телеграм» (Telegram), наглядно показано взаимодействие с пользователем. Определены и обобщены различия входных параметров и получаемых в результате работы программы конкретных нуклеотидных структур. Также в сопоставлении даны временные затраты генерации последовательностей при различных входных данных. Изучены короткие последовательности, различающиеся по типу, длине, GC-составу и содержанию динуклеотидов. В табличном виде показано, как в этом случае соотносятся входные и выходные параметры.
Результаты исследования. Созданное программное обеспечение сравнили с существующими генераторами нуклеотидных последовательностей. Установлено, что генерируемые последовательности отличаются по структуре от известных ДНК-последовательностей живых организмов, а значит, могут быть использованы в качестве вспомогательных или маскирующих олигонуклеотидов, пригодных для молекулярно-биологических манипуляций (например — реакции амплификации), а также для хранения в молекулах ДНК небиологической информации (изображений, текстов и т. д.). Предложенное решение дает возможность формировать специфические последовательности длиной от 20 до 5 000 нуклеотидов с заданным числом динуклеотидов и без гомополимерных участков. Более жесткие условия генерации снимают известные ограничения и позволяют создавать квазислучайные последовательности нуклеотидов по заданным входным параметрам. Кроме количества и длины последовательностей можно заранее определить GC-состав, содержание динуклеотидов и природу нуклеиновой кислоты (ДНК или РНК).

Приводятся примеры коротких последовательностей, различающихся по длине, GC-составу и содержанию динуклеотидов.
Полученные 30-нуклеотидные последовательности прошли проверку. Установлено отсутствие 100-процентной гомологии с известными ДНК-последовательностями живых организмов. Максимальное совпадение наблюдалось для сгенерированных последовательностей длиной 25 нуклеотидов (сходство около 80 %). Таким образом доказано, что GATCGGenerator может с высокой эффективностью генерировать небиологические нуклеотидные последовательности.
Обсуждение и заключение. Новый генератор позволяет создавать нуклеотидные последовательности *in silico* с заданным GC-составом. Решение дает возможность исключить гомополимерные фрагменты, что качественно улучшает физико-химическую стабильность последовательностей.

***Ключевые слова*:** GATCGGenerator, генератор нуклеотидных последовательностей, синтетические нуклеиновые кислоты, случайные последовательности, хранение данных в ДНК, стеганография, NYRN-олигонуклеотиды, вычисления с помощью ДНК, криптография, ДНК-метчики в гидрологии

Благодарности: авторы выражают признательность рецензентам за ценные замечания, способствовавшие улучшению статьи.

Финансирование. Работа выполнена в рамках гранта РФФИ 20–07–00222.

Для цитирования. Кирьянова О.Ю., Гарафутдинов Р.Р., Губайдуллин И.М. Чемерис А.В. GATCGGenerator: новый генератор для создания квазислучайных нуклеотидных последовательностей. *Advanced Engineering Research (Rostov-on-Don).* 2023;23(3):296–306. https://doi.org/10.23947/2687-1653-2023-23-3-296-306

Original article

GATCGGenerator: New Software for Generation of Quasirandom Nucleotide Sequences

Olga Yu. Kiryanova[1] ✉, **Ravil R. Garafutdinov**[2], **Irek M. Gubaydullin**[3], **Aleksei V. Chemeris**[2]
[1] Ufa State Petroleum Technological University, Ufa, Russian Federation
[2] Institute of Biochemistry and Genetics, Ufa Federal Research Center, RAS, Ufa, Russian Federation
[3] Institute of Petrochemistry and Catalysis, RAS, Ufa, Russian Federation
✉ olga.kiryanova27@gmail.com

Abstract
Introduction. In recent decades, knowledge about DNA has been increasingly used to solve biological problems (calculations using DNA, long-term storage of information). Principally, we are talking about cases when it is required to select artificial nucleotide sequences. Special programs are used to create them. However, existing generators do not take into account the physicochemical properties of DNA and do not allow obtaining sequences with a pronounced “non-biological” structure. In fact, they generate sequences by distributing nucleotides randomly. The objective of this work is to create a generator of quasirandom sequences with a special nucleotide structure. It should take into account some physicochemical features of nucleotide structures, and it will be involved in storing non-biological information in DNA.
Materials and Methods. A new GATCGGenerator software for generating quasirandom sequences of nucleotides was described. It was presented as SaaS (from “software as a service”), which provided its availability from various devices and platforms. The program generated sequences of a certain structure taking into account the guanine-cytosine (GC) composition and the content of dinucleotides. The performance of the new program algorithm was presented. The requirements for the generated nucleotide sequences were set using a chat in Telegram, the interaction with the user was clearly shown. The differences between the input parameters and the specific nucleotide structures obtained as a result of the program were determined and generalized. Also, the time costs of generating sequences for different input data were given in comparison. Short sequences differing in type, length, GC composition and dinucleotide content were studied. The tabular form shows how the input and output parameters are correlated in this case.
Results. The developed software was compared to existing nucleotide sequence generators. It has been established that the generated sequences differ in structure from the known DNA sequences of living organisms, which means that they can be used as auxiliary or masking oligonucleotides suitable for molecular biological manipulations (e.g., amplification reactions), as well as for storing non-biological information (images, texts, etc.) in DNA molecules. The proposed solution makes it possible to form specific sequences from 20 to 5 000 nucleotides long with a given number of dinucleotides and without homopolymer fragments. More stringent generation conditions remove known limitations and provide the creation of quasirandom sequences of nucleotides according to specified input parameters. In addition to the number and

length of sequences, it is possible to determine the GC composition, the content of dinucleotides, and the nature of the nucleic acid (DNA or RNA) in advance. Examples of short sequences differing in length, GC composition and dinucleotide content are given. The obtained 30-nucleotide sequences were tested. The absence of 100 % homology with known DNA sequences of living organisms was established. The maximum coincidence was observed for the generated sequences with a length of 25 nucleotides (similarity of about 80 %). Thus, it has been proved that GATCGGenerator can generate non-biological nucleotide sequences with high efficiency.
Discussion and Conclusion. The new generator provides the creation of nucleotide sequences *in silico* with a given GC composition. The solution makes it possible to exclude homopolymer fragments, which improves qualitatively the physicochemical stability of sequences.

Keywords: GATCGGenerator, nucleotide sequences generator, synthetic nucleic acids, random sequences, data storage in DNA, steganography, NYRN-oligonucleotides, calculations with DNA, cryptography, DNA-tagging in hydrology

Acknowledgements: the authors would like to thank the reviewers for valuable comments that contributed to the improvement of the article.

Funding information. The research is done on RFFI grant no. 20‒07‒00222.

For citation. Kiryanova OYu**,** Garafutdinov RR, Gubaydullin IM, Chemeris AV**.** GATCGGenerator: New Software for Generation of Quasirandom Nucleotide Sequences. *Advanced Engineering Research (Rostov-on-Don).* 2023;23(3): 296–306. https://doi.org/10.23947/2687-1653-2023-23-3-296-306

Введение. ДНК является уникальным биополимером, обеспечивающим хранение, передачу и воспроизведение генетической информации в живых организмах. Молекулы ДНК состоят из четырех типов нуклеотидов, содержащих азотистые основания аденин (A), гуанин (G), цитозин (C), тимин (T). Их возможные комбинации обеспечивают нуклеотидные последовательности, формирующие функциональные генетические элементы. В молекулярной биологии и генетике основные работы ведутся с нуклеотидными последовательностями живых организмов, однако возрастает потребность в создании искусственных последовательностей, особенно при решении небиологических задач (например, ДНК-вычисления [1, 2], хранение в ДНК [3], криптография [4], ДНК-метчики в гидрологии [5] и др.).

Как ожидается, к концу 2040 года объем информации достигнет нескольких йоттабайт (10^{24}), что требует ее структурирования и хранения. Оба этих процесса существенно влияют на потребление энергетических ресурсов, а также на производство устройств хранения данных и периферийных устройств (жесткие диски, твердотельные накопители). Для хранения такого количества информации требуется более 10^9 кг особо чистого кремния [6], которого может не хватить. Решение видится в использовании принципов ДНК для работы с масштабными объемами данных.

Нуклеотидные последовательности легко оцифровываются путем присвоения соответствующих двоичных кодов отдельным нуклеотидам [7–11] или блокам нуклеотидов [12–14], поэтому текстовые, графические или мультимедийные файлы можно преобразовывать в последовательности нуклеотидов [15–18]. Искусственные нуклеотидные последовательности можно составить вручную или сгенерировать с помощью специального программного обеспечения (генераторы ДНК) в зависимости от решаемых задач. Некоторые генераторы ДНК разрабатывались как самостоятельные приложения, другие — как часть программных пакетов, предназначенных для решения общих [19] [1, 2, 3, 4, 5] или специфических задач [20]. Как правило, генераторы ДНК разработаны на основе комбинаторных подходов и производят случайные последовательности заданной длины гуанин-цитозинового (GC) состава. Однако такие программные решения не учитывают химические свойства нуклеотидов и не позволяют получать последовательности с определенной структурой (например, без гомополимерных участков или длинных повторяющихся мотивов). Поэтому создаваемые такими генераторами последовательности не всегда можно воспроизвести в лабораторных условиях. Кроме того, такие последовательности могут быть идентичны существующим в природе фрагментам ДНК, что вносит неоднозначность при попытках закодировать информацию небиологического характера.

[1] Nucleotide Sequence Generator // nucleotide-generator.herokuapp.com : [сайт]. URL: https://nucleotide-generator.herokuapp.com/ (дата обращения: 01.12.2022).
[2] DNA Sequence Tools: Random Sequence Generator // molbiotools.com : [сайт]. URL: http://www.molbiotools.com/randomsequencegenerator.html (дата обращения: 01.12.2022).
[3] Random DNA Sequence Generator // faculty.ucr.edu : URL: http://www.faculty.ucr.edu/~mmaduro/random.htm (дата обращения: 02.12.2022).
[4] Random DNA Sequence GenScript // genscript.com : [сайт]. URL: https://www.genscript.com/sms2/random_dna.html (дата обращения: 04.12.2022).
[5] Random DNA Generator // Computer software : [сайт]. URL: http://54.235.254.95/cgi-bin/gd/gdRandDNA.cgi (дата обращения: 04.12.2022).

Цель представленной работы — создание генератора нуклеотидных последовательностей особой структуры, которые можно применять при кодировании текстовой, графической и другой информации в молекулах ДНК.

Материалы и методы. Определены критерии, которые следует иметь в виду при создании последовательностей. Учтена необходимость варьировать GC-состав, задавать определенное количество динуклеотидов, исключить гомополимерные участки в последовательностях.

Коллектив авторов разработал программу GATCGGenerator на языке Python 3.6 (Anaconda distribution)[6]. Для создания бота[7] в «Телеграм» (Telegram) использовали Numpy 1.19 [21] и библиотеку Python GATCGGenerator. Решение предоставляется как SaaS (от англ. software as a service — программное обеспечение как услуга), что открывает возможность доступа с разных устройств и платформ.

Входные параметры: количество последовательностей, их длина, GC-состав и содержание динуклеотидов. Генератор исключает повторы длиной от двух нуклеотидов более четырех раз. Результат представлен в виде файла CSV, который содержит следующую информацию: последовательность, GC-состав и количество всех нуклеотидов.

Повторы и гомополимерные фрагменты хранятся в виде отдельного списка. Сначала случайным образом генерируется последовательность из четырех элементов (random.choice(nuc), где nuc = 'ACGT'). Затем выполняется поиск повторов. Если встречается хотя бы один элемент из списка, выполняется новая случайная генерация. Далее рассчитывается GC- и NN-состав. Если NN-состав не соответствует заданному пользователем диапазону, парный нуклеотид заменяется случайным образом и пересчитывается GC-состав. Если последовательность соответствует входным параметрам, она записывается во множество последовательностей (sequences).

Ниже представлена работа алгоритма программы.

Type — тип; GCmin, GCmax — диапазон возможного содержания GC; NNmin, NNmax — диапазон возможного содержания динуклеотидов NN%; N — количество; S — последовательность; l — длина последовательности; count — общее количество последовательностей

Псевдокод

```
Начало
Ввод (Type, GC, NN, N)
Генерация списка повторяющихся мотивов, гомополимерных участков rep.list
Count = 0
sequences = set()
IF i ≤ N?
    Шаг 1. S = random.choice(‘AGCT’)
    IF (rep.list(k) ⊂ S?)
        Возврат на шаг 1.
    ELSE
        NN = len(DI_REGEX.findall(''.join(S)))
            NN_perc = (NN × 2 / l) × 100
    IF NNmin ≤NN_perc ≤ NNmax
        GC = S.count('G') + S.count('C') / l ×100
    IF GCmin ≤ GC ≤ GCmax
    IF type == DNA
    Шаг 2.
                                A_perc = S.count('A') / l × 100
                                    G_perc = S.count('G') / l × 100
                                    C_perc = S.count('C') / l × 100
                                    T_perc = S.count('T') / l × 100
                                U_perc = S.count('U') / l × 100
    Count = count +1
                          sequences.add(S)
      ELSE S = S.replace('T', 'U')
          Шаг 2.)
                      ELSE
    Возврат на шаг 1.
```

[6] Anaconda / Anaconda Inc. // anaconda.com : [сайт]. URL: https://www.anaconda.com/ (дата обращения: 20.01.2023).

[7] Python telegram bot // github.com : [сайт]. URL: https://github.com/python-telegram-bot/python-telegram-bot (дата обращения: 01.12.2022).

ELSE
Случайная замена второго повторяющегося символа,
GC = S.count('G') + S.count('C') / l ×100
Вывод Sequences: (S, GC%, NN%, A%, G%, C%, T/U%)
Конец

Требования к сгенерированным нуклеотидным последовательностям задаются с помощью чата в Telegram. Пример взаимодействия с пользователем показан на рис. 1.

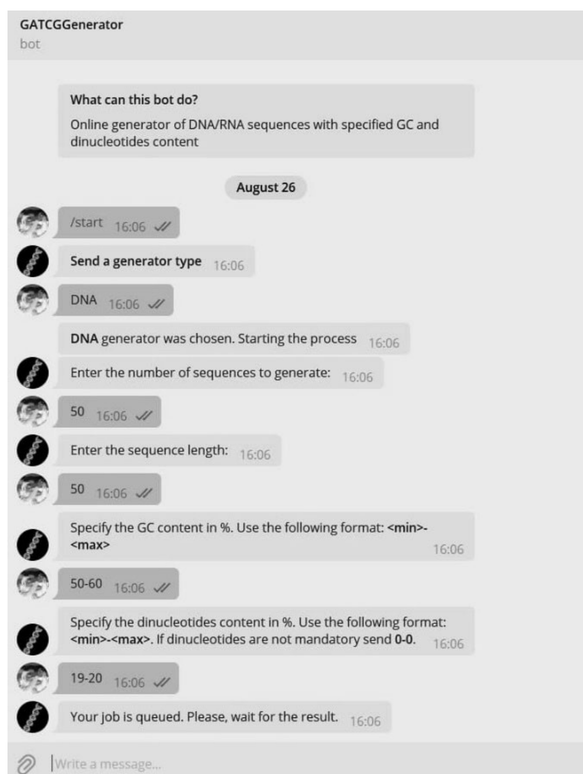


Рис. 1. Пример пользовательского чата в Telegram

В рамках представленной работы сравнивались функциональные возможности генераторов случайных последовательностей и GATCGGenerator. Определялись различия входных параметров и получаемых в результате работы программы конкретных нуклеотидных структур (таблица 1).

Таблица 1

Сравнение функциональных возможностей GATCGGenerator
с другими генераторами нуклеотидных последовательностей

<table>
<tr><th></th><th>GATCGGenerator [20]</th><th>Nucleotide Sequence Generator[8]</th><th>DNA Sequence Tools: Random Sequence Generator[9]</th><th>Random DNA Sequence Generator[10]</th><th>Random DNA Sequence[11]</th><th>Random DNA Generator[12]</th></tr>
<tr><td>Максимальная длина (нуклеотиды)</td><td>5 000</td><td colspan="3">1 000 000</td><td>10 000</td><td>1 000</td></tr>
<tr><td>Число последовательностей</td><td>100</td><td colspan="3">1</td><td>1; 10; 50; 100</td><td>100</td></tr>
<tr><td>Ввод GC-состава (%)</td><td>+</td><td rowspan="2">–</td><td colspan="2">+</td><td>–</td><td>+ (*)</td></tr>
<tr><td>GC- состав (%)</td><td rowspan="2">интервал</td><td colspan="2">число</td><td>–</td><td>число</td></tr>
<tr><td>Ввод NN-состава (%)</td><td colspan="5" rowspan="2">–</td></tr>
<tr><td>Отсутствие гомополимерных участков</td><td>+</td></tr>
<tr><td>Тип последовательности</td><td>ДНК/РНК</td><td>ДНК</td><td>ДНК/РНК/ Протеин</td><td colspan="3">ДНК</td></tr>
<tr><td>Вывод результатов</td><td>.CSV file</td><td colspan="5">Текст на экране</td></tr>
<tr><td colspan="7">(*) Пользователь вводит АТ-состав</td></tr>
</table>

GATCGGenerator обладает более широким функционалом, дает возможность пользователю указывать количество динуклеотидов, создавать последовательности без протяженных гомополимерных участков и повторов, влияющих на успешность эксперимента. В существующих генераторах возможно только варьирование GC-состава.

Программа, созданная авторами данной научной работы, генерирует заданное количество квазислучайных последовательностей нуклеотидов, не имеющих гомологии с природной ДНК, но пригодных для молекулярно-биологических манипуляций.

Результаты исследования. GATCGGenerator позволяет генерировать специфические последовательности ДНК или РНК длиной от 20 до 5 000 нуклеотидов, содержащие заданное количество динуклеотидов и не содержащие гомополимерных участков (не более двух одинаковых нуклеотидов, расположенных рядом). Более жесткие условия генерации могут привести к длительному подбору последовательностей. В качестве примера приведем небольшой диапазон возможного содержания гуанина и цитозина и динуклеотидов (допустим, GC-состав 45–50 % и NN-состав 10–20 %). Продолжительность работы программы для различных входных данных представлена в таблице 2.

Таблица 2

Временные затраты генерации последовательностей при различных входных данных

Входные данные				Время, с
Длина	Число	GC, %	NN, %	
20	10	50–60	20–50	3,45
30	10	50–60	20–50	3,91
20	10	50–60	40–50	9,74
30	10	50–60	40–50	9,53
30	10	40–50	20–20	8,80
1 000	100	45–50	40–50	11,49
2 000	100	45–50	10–20	240,25
5 000	100	50–60	20–50	11,57

GATCGGenerator благодаря более жестким условиям генерации последовательностей снимает ограничения известных генераторов ДНК и создает квазислучайные последовательности нуклеотидов в зависимости от заданных входных параметров. Можно указать необходимое количество последовательностей, их длину, GC-состав и содержание динуклеотидов, а также природу нуклеиновой кислоты (ДНК или РНК). Например, созданные с помощью GATCGGenerator последовательности могут быть использованы в ДНК-стеганографии, применяемой для защиты и передачи информации путем сокрытия содержания сообщения в последовательности нуклеотидов [3].

Предлагаемое программное решение (GATCGGenerator) позволяет получать набор квазислучайных последовательностей нуклеотидов в зависимости от заданных пользователем входных параметров (тип нуклеиновой кислоты, длина последовательности, GC- и динуклеотидный состав). GATCGGenerator исключает наличие любых нуклеотидных повторов и гомополимерных участков длиннее трех элементов. Сгенерированные последовательности могут быть использованы как служебные или маскирующие (например, в ДНК-стеганографии) и подходят для любых небиологических ферментативных манипуляций. Можно сгенерировать множество искусственных нуклеотидных последовательностей и использовать их для создания универсальной олиготеки, пригодной для многократного кодирования небиологических данных и их длительного хранения.

Данные, представленные в таблице 3, обобщенно демонстрируют результаты работы программы. Для определенного типа нуклеиновой кислоты (в данном случае ДНК) показаны: содержание динуклеотидов (NN %), количество сгенерированных последовательностей, их длина (нуклеотиды — нт) и GC-состав.

Полученные 30-нуклеотидные последовательности проверили с помощью инструмента Blast от NCBI. Выявлено отсутствие 100-процентной гомологии с известными ДНК-последовательностями живых организмов. Максимальное совпадение наблюдалось для сгенерированных последовательностей длиной 25 нуклеотидов (сходство около 80 %). Это свидетельствует о способности GATCGGenerator с высокой эффективностью генерировать небиологические нуклеотидные последовательности. Можно считать, что сгенерированные таким образом последовательности не имеют абсолютного совпадения с нуклеотидными фрагментами живых организмов.

Таблица 3

Примеры коротких последовательностей, различающихся по длине, GC-составу и содержанию динуклеотидов, %

Входные параметры					Нуклеотидная последовательность, 5’→3’	Выходные параметры		
Тип	Число	Длина, нт	GC, %	NN, %*		Длина, нт	GC, %	NN, %*
ДНК	5	30	41–50	20	CT**GG****TATATC**GGAA**TCATATCGCGCAGTGT	30	46,7	20,0
					AATCAGCTAGTA**GG**ACGCAGTAGTG**AA**TCA	30	43,3	20,0
					G**AA**TGTAGT**CC**TA**GG**CACATACTACGTAGC	30	46,7	20,0
					AG**TT**GCACTG**AA**GTCTATGATCT**GG**CATGC	30	46,7	20,0
					GACACACTACTAT**GG**ACGTGA**GG**CAC**TT**AC	30	50,0	20,0
	5		51–60		TCAGCTCAGCG**CCAA**TCGAGC**TT**ATAGTGC	30	53,3	20,0
					GA**GG**CTATCGTC**AA**GCATAGA**CC**GTGTGCT	30	53,3	20,0
					GACTCAGTAGCTGCT**CCGG**ACATACAG**CC**T	30	56,7	20,0
					TCGCGCG**TT**AGAC**TT**A**GG**TCTCATCGCAGC	30	56,7	20,0
					ACGCTCACA**GG**AG**TT**CGCATCG**AA**CGATGC	30	56,7	20,0
	5		41–50	0	ACGACAGTGATATAGCACGACGTGCTCATA	30	46,7	0,0
					GACTACATCTGATAGTACACGTGCTGCACT	30	46,7	0,0
					TCTATCTCTGCTAGAGCGCTCGTCACTCTA	30	50,0	0,0
					TCTGATCTACTATAGCGATACGTGAGAGTG	30	43,3	0,0
					ACACATATATCGACGCACGCGTCGTAGTAC	30	50,0	0,0
	5	50	41–60	20	TGCATGA**CC**ATGCTTGC**GG**TAGACA**TT**CAGA CGCGCG**AA**TAGTA**GG**ACGA	50	52,0	20,0
					GCATACGAGT**GG**CATACATA**TT**AGACTATAC **GG**TAGTGCATAT**GG**TGC**AA**	50	42,0	20,0
					CTGAGACT**CC**TCTCTGT**GG**AGCT**CC**TAGTA**C** **C**GTCACGCGTGCTCTG**AA**G	50	58,0	20,0
					CTGTGTG**AA**CATACGATGCA**TT**CTCATCTC**GG** TAT**GG**CTG**AA**GTGCACAT	50	46,0	20,0
					GCGCTGACGTCAT**GGTT**CATA**CCAA**TGTAGC ATGATGTGCGATA**GG**CACA	50	50,0	20,0

*NN показывает долю (%) содержащихся динуклеотидов в нуклеотидной последовательности.
**Динуклеотиды выделены жирным шрифтом.

В этом случае в качестве удобного носителя информации можно задействовать специальные ДНК-олигонуклеотиды искусственного происхождения, содержащие информативную и служебную части. Недавно авторы данной работы предложили использовать NYRN-олигонуклеотиды [14], состоящие из:

– внутренней части (YR)n, кодирующей зашифрованную информацию;

– служебных (вспомогательных) частей S1 и S2, фланкирующих последовательность (YR)n (рис. 2).

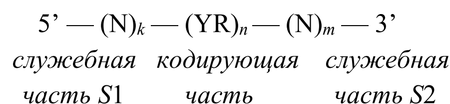


Рис. 2. Структура NYRN-олигонуклеотидов: N — вырожденные нуклеотиды; Y — пиримидины (C или T); R — пурины (A или G); *k*, *n*, *m* — индексы, соответствующие длине части

Длина участков (*n*, *k* и *m*) может варьироваться, но структура служебных частей должна обеспечивать успешное протекание реакций амплификации (длина более 18 нт, 40–60 % GC-состав, отсутствие гомополимерных участков и повторов). GATCGGenerator позволяет включать динуклеотиды NN, содержащие одинаковые парные нуклеотиды (например, AA, GG, CC, TT или UU для РНК), которые могут повысить специфичность молекулярной гибридизации нуклеиновых кислот.

Обсуждение и заключение. Итак, по итогам выполненной научной работы предложено программное решение (GATCGGenerator), которое в сравнении с традиционными подходами предполагает более жесткие

условия генерации последовательностей. Благодаря этой его особенности снимаются ограничения известных генераторов ДНК и формируются квазислучайные последовательности нуклеотидов в зависимости от заданных входных параметров. Исследованы полученные 30-нуклеотидные последовательности. Проверка позволила установить отсутствие 100-процентной гомологии с известными ДНК-последовательностями живых организмов. Максимально (примерно на 80 %) совпали сгенерированные последовательности длиной 25 нуклеотидов.

Отметим также, что для сокрытия информации в NYRN-олигонуклеотидах, необходимо их смешать с маскирующей ДНК. Маскирующие последовательности должны быть аналогичны последовательностям NYRN-олигонуклеотидов, чтобы при попытке считывания скрытой информации невозможно было их распознать без ключевых последовательностей. Адресату должны быть известны ключевые последовательности — праймеры к служебным участкам NYRN-олигонуклеотидов. Адресат может расшифровать переданное сообщение путем выделения информативных последовательностей нуклеотидов с помощью полимеразной цепной реакции с последующим секвенированием. Набор NYRN- и маскирующих олигонуклеотидов можно легко получить с помощью GATCGGenerator, синтезировать, а затем сохранить в виде олиготеки. Для этого достаточно определить оптимальные NYRN-олигонуклеотиды с последующим заполнением олиготеки. В дальнейшем планируется проведение лабораторных экспериментов с целью апробации предложенного метода хранения небиологической информации и проверки жизнеспособности олиготек, получаемых с помощью генератора.

Список литературы

1. Малинецкий Г.Г., Митин Н.А., Науменко С.А. Нанобиология и синергетика. Проблемы и идеи. *Препринты Института прикладной математики им. М.В. Келдыша РАН.* 2005;29:1–26. URL: http://mi.mathnet.ru/ipmp722 (дата обращения: 01.06.2023).

2. Katz E. (ed) *DNA- and RNA-Based Computing Systems*, 1st ed. Weinheim: Wiley-VCH; 2021. 408 p.

3. Ceze L., Nivala J., Strauss K. Molecular Digital Data Storage Using DNA. *Nature Reviews Genetics*. 2019;20:456–466. https://doi.org/10.1038/s41576-019-0125-3

4. Kaundal A.K., Verma A.K. DNA Based Cryptography: A Review. *International Journal of Information and Computation Technology*. 2014;4(7):693–698.

5. Aquilanti L., Clementi F., Landolfo S., Nanni T., Palpacelli S., Tazioli A. A DNA Tracer Used in Column Tests for Hydrogeology Applications. *Environmental Earth Sciences*. 2013;70:3143–3154. https://doi.org/10.1007/s12665-013-2379-y

6. Zhirnov V., Zadegan R.M., Sandhu G.S., Church G.M., Hughes W. Nucleic Acid Memory. *Nature Materials*. 2016;15:366–370. https://doi.org/10.1038/nmat4594

7. Yetisen A.K., Davis J., Coskun A.F., Church G.M., Seok Hyun Yun. Bioart. *Trends in Biotechnology*. 2015;33(12):724–734. https://doi.org/10.1016/j.tibtech.2015.09.011

8. Dokyun Na. DNA Steganography: Hiding Undetectable Secret Messages within the Single Nucleotide Polymorphisms of a Genome and Detecting Mutation-Induced Errors. *Microbial Cell Factories*. 2020;19(128):1–9. https://doi.org/10.1186/s12934-020-01387-0

9. Shuhong Jiao, Goutte R. Code for Encryption Hiding Data into Genomic DNA of Living Organisms. In: *Proc. 9th International Conference on Signal Processing*. Beijing: IEEE; 2008. P. 2166–2169. https://doi.org/10.1109/ICOSP.2008.4697576

10. Masanori Arita. Writing Information into DNA. In book: N. Jonoska, G. Păun, G. Rozenberg (eds). *Aspects of Molecular Computing. Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer; 2004. P. 23–35. https://doi.org/10.1007/978-3-540-24635-0_2

11. Church G.M., Yuan Gao, Sriram Kosuri. Next-Generation Digital Information Storage in DNA. *Science*. 2012;337(6102):1628. https://doi.org/10.1126/science.1226355

12. K.A. Schouhamer Immink, Kui Cai. Design of Capacity-Approaching Constrained Codes for DNA Based Storage Systems. *IEEE Communications Letters*. 2018;22(2):224–227. https://doi.org/10.1109/LCOMM.2017.2775608

13. Nozomu Yachie, Kazuhide Sekiyama, Junichi Sugahara, Yoshiaki Ohashi, Masaru Tomita. Alignment-Based Approach for Durable Data Storage into Living Organisms. *Biotechnology Progress*. 2007;23(2):501–505. https://doi.org/10.1021/bp060261y

14. Garafutdinov R.R., Sakhabutdinova A.R., Slominsky P.A. Aminev F.G., Chemeris A.V. A New Digital Approach to SNP Encoding for DNA Identification. *Forensic Science International*. 2020;317:110520. https://doi.org/10.1016/j.forsciint.2020.110520

15. Ailenberg M., Rotstein O.D. An Improved Huffman Coding Method for Archiving Text, Images, and Music Characters in DNA. *BioTechniques*. 2009;47(3):747–754. https://doi.org/10.2144/000113218

16. Doricchi A., Platnich C.M., Gimpel A., Horn F., Earle M., Lanzavecchia G., et al. Emerging Approaches to DNA Data Storage: Challenges and Prospects. *ACS Nano*. 2022;16(11):17552–17571. https://doi.org/10.1021/acsnano.2c06748

17. Sakhabutdinova A.R., Mikhailenko K.I., Garafutdinov R.R., Kiryanova O.Yu., Sagitova M.A., Sagitov A.M., et al. Non-Biological Application of DNA Molecules. *Biomics*. 2019;11(3):344–377. https://doi.org/10.31301/2221-6197.bmcs.2019-28

18. Garafutdinov R.R., Chemeris D.A., Sakhabutdinova A.R. Chemeris A.V., Kiryanova O.Yu., Mikhaylenko C.I. Encoding of Non-Biological Information for its Long-Term Storage in DNA. *Biosystems*. 2022;(215–216):104664. https://doi.org/10.1016/j.biosystems.2022.104664.9

19. Кирьянова О.Ю., Кирьянов И.И., Гарафутдинов Р.Р., Чемерис А.В., Губайдуллин И.М. *GATCGGenerator*. Свидетельство о регистрации программы для ЭВМ № RU 2021667097. 2021.

20. Borzov E.A., Marakhonov A.V., Ivanov M.V., Drozdova P.B., Baranova A.V., Skoblov M.Yu. RANDTRAN: Random Transcriptome Sequence Generator that Accounts for Partition Specific Features in Eukaryotic mRNA Datasets. *Molecular Biology*. 2014;48:749–756. https://doi.org/10.1134/S0026893314050021

21. Harris C.R., Millman K.J., van der Walt S.J., Gommers R., Virtanen P., Cournapeau D., et al. Array Programming with NumPy. *Nature*. 2020;585:357–362. https://doi.org/10.1038/s41586-020-2649-2

References

1. Malinetski GG, Mitin NA, Naumenko SA. Nanobiology and Synergetics. Problems and Ideas. Part 2. *Keldysh Institute Preprints.* 2005;29:1–26. URL: http://mi.mathnet.ru/ipmp722 (accessed: 01.06.2023).

2. Katz E. (ed) *DNA- and RNA-Based Computing Systems*, 1st ed. Weinheim: Wiley-VCH; 2021. 408 p.

3. Ceze L, Nivala J, Strauss K. Molecular Digital Data Storage Using DNA. *Nature Reviews Genetics*. 2019;20:456–466. https://doi.org/10.1038/s41576-019-0125-3

4. Kaundal AK, Verma AK. DNA Based Cryptography: A Review. *International Journal of Information and Computation Technology*. 2014;4(7):693–698.

5. Aquilanti L, Clementi F, Landolfo S, Nanni T, Palpacelli S, Tazioli A. A DNA Tracer Used in Column Tests for Hydrogeology Applications. *Environmental Earth Sciences*. 2013;70:3143–3154. https://doi.org/10.1007/s12665-013-2379-y

6. Zhirnov V, Zadegan RM, Sandhu GS, Church GM, Hughes W. Nucleic Acid Memory. *Nature Materials*. 2016;15:366–370. https://doi.org/10.1038/nmat4594

7. Yetisen AK, Davis J, Coskun AF, Church GM, Seok Hyun Yun. Bioart. *Trends in Biotechnology*. 2015;33(12):724–734. https://doi.org/10.1016/j.tibtech.2015.09.011

8. Na D. DNA Steganography: Hiding Undetectable Secret Messages within the Single Nucleotide Polymorphisms of a Genome and Detecting Mutation-Induced Errors. *Microbial Cell Factories*. 2020;19(128):1–9. https://doi.org/10.1186/s12934-020-01387-0

9. Shuhong Jiao, Goutte R. Code for Encryption Hiding Data into Genomic DNA of Living Organisms. In: *Proc. 9th International Conference on Signal Processing*. Beijing: IEEE; 2008. P. 2166–2169. https://doi.org/10.1109/ICOSP.2008.4697576

10. Masanori Arita. Writing Information into DNA. In book: N. Jonoska, G. Păun, G. Rozenberg (eds). *Aspects of Molecular Computing. Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer; 2004. P. 23–35. https://doi.org/10.1007/978-3-540-24635-0_2

11. Church GM, Yuan Gao, Sriram Kosuri. Next-Generation Digital Information Storage in DNA. *Science*. 2012;337(6102):1628. https://doi.org/10.1126/science.1226355

12. KA Schouhamer Immink, Kui Cai. Design of Capacity-Approaching Constrained Codes for DNA Based Storage Systems. *IEEE Communications Letters*. 2018;22(2):224–227. https://doi.org/10.1109/LCOMM.2017.2775608

13. Nozomu Yachie, Kazuhide Sekiyama, Junichi Sugahara, Yoshiaki Ohashi, Masaru Tomita. Alignment-Based Approach for Durable Data Storage into Living Organisms. *Biotechnology Progress*. 2007;23(2):501–505. https://doi.org/10.1021/bp060261y

14. Garafutdinov RR, Sakhabutdinova AR, Slominsky PA, Aminev FG, Chemeris AV. A New Digital Approach to SNP Encoding for DNA Identification. *Forensic Science International*. 2020;317:110520. https://doi.org/10.1016/j.forsciint.2020.110520

15. Ailenberg M, Rotstein OD. An Improved Huffman Coding Method for Archiving Text, Images, and Music Characters in DNA. *BioTechniques*. 2009;47(3):747–754. https://doi.org/10.2144/000113218

16. Doricchi A, Platnich CM, Gimpel A, Horn F, Earle M, Lanzavecchia G, et al. Emerging Approaches to DNA Data Storage: Challenges and Prospects. *ACS Nano*. 2022;16(11):17552–17571. https://doi.org/10.1021/acsnano.2c06748

17. Sakhabutdinova AR, Mikhailenko KI, Garafutdinov RR, Kiryanova OYu, Sagitova MA, Sagitov AM, et al. Non-Biological Application of DNA Molecules. *Biomics*. 2019;11(3):344–377. https://doi.org/10.31301/2221-6197.bmcs.2019-28

18. Garafutdinov RR, Chemeris DA, Sakhabutdinova AR, Chemeris AV, Kiryanova OYu, Mikhaylenko CI. Encoding of Non-Biological Information for its Long-Term Storage in DNA. *Biosystems*. 2022;(215–216):104664. https://doi.org/10.1016/j.biosystems.2022.104664.9

19. Kiryanova OYu, Kiryanova II, Garafutdinov RR, Chemeris DA, Gubaidullin IM. *GATCGGenerator*. Certificate of Software Registration No. RU 2021667097. 2021. (In Russ.)

20. Borzov EA, Marakhonov AV, Ivanov MV, Drozdova PB, Baranova AV, Skoblov MYu. RANDTRAN: Random Transcriptome Sequence Generator that Accounts for Partition Specific Features in Eukaryotic mRNA Datasets. *Molecular Biology*. 2014;48:749–756. https://doi.org/10.1134/S0026893314050021

21. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array Programming with NumPy. *Nature*. 2020;585:357–362. https://doi.org/10.1038/s41586-020-2649-2

Поступила в редакцию 07.06.2023
Поступила после рецензирования 29.06.2023
Принята к публикации 03.07.2023

Об авторах:

Ольга Юрьевна Кирьянова, ассистент кафедры цифровых технологий и моделирования Уфимского государственного нефтяного технического университета (РФ, 450064, Уфа, ул. Космонавтов, 1), Researcher ID, ScopusID, ORCID, AuthorID, olga.kiryanova27@gmail.com

Равиль Ринатович Гарафутдинов, кандидат химических наук, заведующий лабораторией физико-химических методов анализа биополимеров Института биохимии и генетики — обособленного структурного подразделения Федерального государственного бюджетного научного учреждения «Уфимский федеральный исследовательский центр» (РФ, 450054, Уфа, пр. Октября, 71), ScopusID, ORCID, AuthorID, garafutdinovr@mail.ru

Ирек Марсович Губайдуллин, доктор физико-математических наук, профессор, заведующий лабораторией математической химии Института нефтехимии и катализа — обособленного структурного подразделения Федерального государственного бюджетного научного учреждения «Уфимский федеральный исследовательский центр» (РФ, 450075, Уфа, пр. Октября, 141), ScopusID, ORCID, AuthorID, irekmars@mail.ru

Алексей Викторович Чемерис, доктор биологических наук, профессор, главный научный сотрудник Института биохимии и генетики — обособленного структурного подразделения Федерального государственного бюджетного научного учреждения «Уфимский федеральный исследовательский центр» (РФ, 450054, Уфа, пр. Октября, 71), ScopusID, ORCID, AuthorID, chemeris@anrb.ru

Заявленный вклад соавторов:

О.Ю. Кирьянова — разработка программного обеспечения, подготовка текста, расчеты, формулировка выводов.

Р.Р. Гарафутдинов — консультирование по предметной области, тестирование ПО, доработка текста, корректировка выводов.

И.М. Губайдуллин — научное руководство, корректировка выводов, доработка текста статьи.

А.В. Чемерис — формирование основной концепции, целей и задач исследования, анализ результатов исследования, доработка текста, корректировка выводов.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Все авторы прочитали и одобрили окончательный вариант рукописи.

Received 07.06.2023
Revised 29.06.2023
Accepted 03.07.2023

About the Authors:

Olga Yu. Kiryanova, Teaching assistant of the Department of Digital Technologies and Modeling, Ufa State Aviation Technical University (1, Kosmonavtov St., Ufa, 450064, RF), Researcher ID, ScopusID, ORCID, AuthorID, olga.kiryanova27@gmail.com

Ravil R. Garafutdinov, Cand.Sci. (Chemistry), Head of the Laboratory of Physico-Chemical Methods of Analysis of Biopolymer, Institute of Biochemistry and Genetics — a separate structural subdivision of the Ufa Federal Research Centre, RAS (71, Oktyabrya Av., Ufa, 450054, Bashkortostan Rep., RF), ScopusID, ORCID, AuthorID, garafutdinovr@mail.ru

Irek M. Gubaydullin, Dr.Sci. (Phys.-Math.), Professor, Head of the Laboratory of Mathematical Chemistry, Institute of Petrochemistry and Catalysis — a separate structural subdivision of the Ufa Federal Research Centre, RAS (141, Oktyabrya Av., Ufa, 450075, Bashkortostan Rep., RF), ScopusID, ORCID, AuthorID, irekmars@mail.ru

Aleksei V. Chemeris, Dr.Sci. (Biology), Professor, Chief Research Fellow, Institute of Biochemistry and Genetics — a separate structural subdivision of the Ufa Federal Research Centre, RAS (71, Oktyabrya Av., Ufa, 450054, Bashkortostan Rep., RF), ScopusID, ORCID, AuthorID, chemeris@anrb.ru

Claimed Contributorship:

OYu Kiryanova: software development, text preparation, calculations, formulation of conclusions.

RR Garafutdinov: consulting on the subject area, software testing, revision of the text, correction of the conclusions.

IM Gubaydullin: academic advising, correction of the conclusions, revision of the text.

AV Chemeris: formulation of the basic concept, research objectives and tasks; analysis of the research results, revision of the text, correction of the conclusions.

Conflict of interest statement: the authors do not have any conflict of interest.

All authors have read and approved the final manuscript.